Accepted for Nucleic Acids Research Jan 25th 2009

Integration of Phenotypic Metadata and Protein Similarity in Archaea Using a Spectral Bipartitioning Approach

Sean D. Hooper[1*], Iain J Anderson[1], Amrita Pati[1], Daniel Dalevi[2], Konstantinos Mavromatis[1], Nikos C Kyrpides[1]

[1] Department of Energy Joint Genome Institute (DOE-JGI), Genome Biology Program, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

[2] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road,
Berkeley, CA 94720, USA

* Corresponding Author

Abstract

In order to simplify and meaningfully categorize large sets of protein sequence data, it is commonplace to cluster proteins based on the similarity of those sequences. However, it quickly becomes clear that the sequence flexibility allowed a given protein varies significantly among different protein families. The degree to which sequences are conserved not only differs for each protein family, but also is affected by the phylogenetic divergence of the source organisms. Clustering techniques that use similarity thresholds for protein families do not always allow for these variations and thus cannot be confidently used for applications such as automated annotation and phylogenetic profiling.

In this work, we applied a spectral bipartitioning technique to all proteins from 53 archaeal genomes. Comparisons between different taxonomic levels allowed us to study the effects of phylogenetic distances on cluster structure. Likewise, by associating functional annotations and phenotypic metadata with each protein, we could compare our protein similarity clusters with both protein function and associated phenotype. Our clusters can be analyzed graphically and interactively online.

Introduction

Clustering is a commonly used method for partitioning large datasets into meaningful groups, i.e. to categorize data by the characteristics they share. The distance from each data point in the set to every other data point is expressed by some measure—typically the amino acid sequence when clustering proteins. Similarity of sequence is then often used to infer some commonality among proteins, whether it be a common evolutionary function, shared domain structure or shared function.

Before proteins are clustered into groups of proteins that share sequence characteristics, it is helpful to view the relationship between proteins as a network where proteins are represented as nodes, and the similarity between proteins are represented as edges between nodes. In this work, we used a spectral clustering approach (1,2) to cluster proteins in this network. This method clusters proteins based on the topology of the entire network rather than based on individual similarities between proteins. Since there is no need to apply arbitrary identity thresholds, spectral clustering is highly suitable for the study of families of proteins which are the result of evolutionary processes such as sequence conservation, sequence divergence and duplication(1).

Proteins can also be clustered at different levels; at the most basic level, we could cluster proteins that have only a general similarity, such as for instance a shared domain. These clusters would reflect the distribution of domains among families of proteins. On a deeper level, we could cluster proteins that are orthologous but do not share a function, and thereafter orthologous proteins of the same function. Additionally, we could also separate orthologous proteins by effects of species divergence. For instance, within a small phylogenetic space (e.g., within a genus), a cluster of orthologous proteins that share the same function can be difficult to separate according to phylogeny, while separation could be easier as the phylogenetic distance increases (e.g. across different phyla). The phylogenetic separation is not necessarily the last in order, since different protein families have varying sequence conservation. It is for instance conceivable that some proteins with strongly conserved sequences (e.g. cell machinery) would not be easily separable at the phylum level.

By applying a spectral clustering algorithm which successively divides clusters of proteins into two child clusters (bipartitioning), we here attempt to assemble a hierarchy of clusters which may reflect these various levels of clustering. This approach partitions clusters of proteins until no further topology-based partition is possible. As a result, different clusters may exhibit very different levels of protein identity depending on the degree of sequence conservation, and different clusters may reflect varying aspects of the evolutionary processes involved. Taking the Archaea as an example, the phylogenetic distance separating the two main phyla (the Crenarchaeota and the Euryarchaeota) is

greater than the distance between two families within the Crenarchaeota (e.g., the Thermoproteaceae and the Sulfolobaceae). For a typical group of orthologous proteins, their sequence similarity would be lower between the phyla than between the families. For clustering methods based on similarity thresholds (see (1) for a discussion of methods), the sequence divergence between the phyla may obscure the similarities in the proteins. In contrast, a topology-based clustering method could group all the members of a protein family, such as the phosphokinases, then partition the crenarchaeal phosphokinases from their euryarchaeal counterparts, and perhaps even further partition the phosphokinases in the Thermoproteaceae from those of the Sulfolobaceae.

Using this spectral bipartitioning clustering approach, we created successive protein similarity clusters for the protein sequences from 53 archaeal genomes. We then assessed how well the clusters distinguished by this method correlated with phylogeny (3) and functional annotation. Likewise, by including phenotypic metadata with each protein, such as the metabolism and habitat of its source organism, metadata co-occurrence profiles analogous to the phylogenetic profiles could be generated, thus adding further value to the protein clustering.

We also developed an intuitive, graphical online tool (available at http://coal.jgi-psf.org/) where these *Spectral Bipartitioning Clusters* (SBCs) can be analyzed and explored further.

Results

Clusters generated by spectral bipartitioning

The cluster architecture produced by spectral bipartitioning is inherently hierarchical since the clusters generated at each level are, in turn, bipartitioned so long as they fulfill the specified topological requirements (see Materials and Methods). *Root* clusters are at the top level of the hierarchy and are designated by integers (e.g., **0** or **1** or **2**). Partitioning of root cluster **1**, for example, would yield two subclusters assigned the identifiers **1.0** and **1.1**. Likewise, further partitioning of **1.1** would yield **1.1.0** and **1.1.1**, and so on. Clusters that do not have any subclusters (i.e. clusters that cannot be partitioned further) are termed *leaf* clusters. In some cases, these designations overlap since a root SBC that cannot be partitioned is also a leaf SBC.

From a total number of 122,452 archaeal proteins (4), we selected those proteins with reciprocal best hits in other genomes, resulting in a sequence similarity matrix of 95,803 proteins(4). Thus, there are no SBCs composed of paralogs from a single species or proteins unique to one species, ensuring that correlations between protein similarity, function and metadata focus on proteins that are present in at

least two species. From this matrix we identified 8463 independent root SBCs with 11 ± 61 members. Successive bipartitioning of these root SBCs produced 10,247 additional SBCs for a total of 18,710. These 18,710 SBCs included 13,586 leaf SBCs (7 ± 10 members), some of which are also root SBCs and others which were generated by partitioning.

Using consistency scores to assess correlations

We assessed how well the protein clustering correlates with (1) the functional annotations of the member proteins; (2) the phylogeny of the source organisms; and (3) four lifestyle-related phenotypic traits of the source organisms. The phenotypic metadata for habitat, oxygen preference, temperature range, and metabolism were extracted from the Genomes Online Database (GOLD) (5) for all 53 archaeal genomes (Supplementary Table S1). Consistency scores were calculated for all leaf SBCs (see Materials and Methods). Scores range from 0 to 1. A high score indicates that one character state dominates the cluster; likewise, a low score denotes that the cluster is heterogeneous with respect to that type of data. When considering temperature preference metadata, for example, a cluster with only hyperthermophilic members would be termed fully consistent and would have a consistency score of 1. When calculating consistency scores for the functional annotations, we ignored a cluster if any members lacked a COG assignment (i.e., we set the consistency score to 0).

Figure 1 shows the distribution of consistency scores for leaf SBCs with respect to functional annotations and all four types of phenotypic metadata. Clusters were highly consistent for functional annotation, mostly due to the relaxed similarity requirements for COG assignments. Leaf SBCs also tend to be consistent with respect to temperature preference, oxygen usage, and metabolism, but significantly less so with respect to habitat. However, it is likely that most of the fully consistent SBCs would be composed of proteins from closely related species, therefore reflecting the phylogeny rather than the phenotype of organisms. Some of the most interesting insights stem from those clusters that include members from distant archaea. For example, when investigating aerobicity, we would pay most attention to those few clusters where the member proteins reside in distantly related aerobic organisms, since this may indicate *Lateral Gene Transfer* (LGT) rather than a vertical inheritance.

Additional factors need to be borne in mind when interpreting consistency scores for the four metadata types that are lifestyle-related phenotypes (temperature preference, oxygen usage, metabolism, and habitat). It can be difficult to distinguish the similarities due to phenotype from those resulting from phylogeny. Since, for example, only a small percentage of proteins may be directly associated with thermal adaptation, one would expect that the majority of the consistent clusters found would be due

simply to phylogeny. A leaf SBC could be consistent with respect to oxygen usage because the member proteins are all specifically related to the aerobic lifestyle, or, it could be consistent simply because the member proteins all reside in very closely related organisms, all of which happen to be aerobic. This situation is illustrated by SBCs **1267.0** and **1093**, all members of which are from aerobes. Although both clusters are fully consistent with respect to oxygen usage, the interpretation of their consistency differs. SBC 1267.0 is composed of three closely related Halobacteria, while SBC **1093** includes both thermoprotei (Crenarchaeota) and halobacteria (Euryarchaeota). Although the clustering of SBC **1093** correlates with oxygen usage, that for SBC **1267.0** appears to be an artifact reflecting predominantly inheritance effects.

Correlating clusters with protein functional annotations

Through subsequent partitioning of a root cluster, it is conceivable that the subclusters would at some point correspond to orthologous proteins, sharing a function. To test how well SBCs correlate with functional annotations, we used the Clusters of Orthologous Genes (COG, (6)) and the Archaeal COG (arCOG, (7)) assignments as benchmarks and calculated the consistency within SBCs. The vast majority of both root and leaf SBCs fall into one of two categories: either all members are assigned to the same COG, or one or more members lack COG assignments and thus the SBC is scored as having "zero" COGs (Table 1). Similarly, the arCOGs(7) were created analogously to COGs, but based exclusively based on 41 Archaea. Due to the differences in methodology between SBCs and (ar)COGs, and since SBCs are based on similarity and not necessarily orthology, we do not expect the two methods to completely overlap. Thus, SBCs may, depending on level, be more specific or more general then (ar)COGs. For instance, the root cluster **998** corresponds to the initiation factor 2 subunit family (pfam01008, (8)), while subsequent child clusters distinguish the orthologous clusters COG1184 and COG0182. Further partitioning results in separating COG1184 based on differences in Halobacterial vs. other archaeal versions of the protein.

Since COGs (and to a lesser extent arCOGs) are typically large collections of proteins which may be only distantly related, one would expect leaf SBCs to be subsets of COGs. Although most leaf SBCs were found to be subsets of COGs, some do contain more than one COG. While root clusters might be expected to contain several COGs prior to partitioning, it is more surprising to find leaf clusters with multiple COGs or arCOGs. There are several explanations for this finding. In some instances, although the proteins in a leaf SBC are assigned to different COGS, they are so similar that the algorithm cannot partition the SBC further. This appears to be the case in SBC **1376** which contains two similar COGs: COG0555 (Na+/proline symporter) and COG4149 (Na+/panthothenate symporter). In this case, the

SBC has only one arCOG, so the latter approach consolidates the functional classification and is consistent with protein similarities. Other times, two COGs within a leaf cluster have the same annotation, thus suggesting that these COGs could be merged. For instance, see SBC 113.0.0 where member COGs COG0668 and COG3264 are both annotated as small-conductance mechanosensitive channels. Another situation is illustrated by the finding of both COG1111 and COG1948 in SBC 222.0. Here, the presence of the ERCC4 domain in both helicases (COG1111) and nucleases (COG1948) blurs the distinction between them.

Sometimes, the resolution of the spectral bipartitioning algorithm partitions a single COG into two or more subclusters. For instance, root SBC 623 contains COG1958. Partitioning of that cluster revealed that there are two subclasses of COG1958 (arCOG00998), one which is conserved in both phyla and another which is unique to the Thermoprotei within the Crenarchaeota. These observations suggest that duplication of the original gene in the Thermoprotei was followed by a functional adaptation within the Thermoprotei class.

In summary, the overall agreement between (ar)COGs and SBCs suggest that a spectral bipartitioning approach is able to not only cluster proteins based on effects of shared domains and phylogenetic divergence, but is also able to reconstruct clusters of orthologous proteins.

Correlating leaf clusters with phenotypic metadata

We found that 41%, 39%, and 32% of leaf clusters are fully consistent with respect to thermal preference, oxygen usage, and metabolism, respectively, but only 22% with respect to habitat. One reason for the lower value for habitat is that it is more difficult to classify habitats into non-overlapping groups. Also, a habitat assigned based on where an archaeaon was observed in nature may not be its optimal habitat. Generally, one must also be aware of false positives (see Methods), and the protein function annotation should always be consulted.

Of particular interest are those fully consistent SBCs that contain at least five members and include representatives of different phyla. These cases are quite rare; accounting for only 0.3% to 1.4% of the fully consistent clusters for the various metadata types. When interpreting these clusters, special care must be taken to distinguish between the effects of phenotype and phylogeny. For instance, a protein family unique to the thermoprotei will appear to correlate strongly with thermophilic temperature preference, but most likely this would be due to the recent divergence of these species. To identify protein families truly associated with thermophilic adaptation, one looks for clusters that span large phylogenetic distances, i.e., that have members from different phyla. By restricting further analyses to

such phylogenetically inconsistent clusters, we would expect to identify the clusters with the strongest correlations with phenotype, but would likely miss others. In the following sections, we discuss some specific SBCs that are consistent with respect to a phenotypic metadata type and also are phylogenetically inconsistent at the phylum level, and offer possible interpretations of these observations.

Thermal preference: Most member proteins in these selected SBCs are annotated as *hypothetical;* those with functional annotations include proteins linked to RNA methyltransferase (SBC **8376.1.0.1.1.1.1.0.0.1.1.1.0**). It has been shown(9) that in some organisms react to increased temperatures by increasing methylation of RNA, so it is therefore not surprising to find this SBC consistent with a thermal preference. Furthermore, it has been reported that reverse gyrase is the only protein unique to hyperthermophiles (10,11). Reverse gyrase resides in the root SBC **1731** and its two subclusters: **1731.1** that contains only proteins from hyperthermophilic Crenarchaeota of the class Thermoprotei and **1731.0** that includes proteins from all three archaeal phyla. This finding is consistent with previous observations implicating LGT in the evolutionary history of reverse gyrase (11).

Two COGs (COG1857 and COG1688) are key components of a novel putative DNA repair system reported to be essential for thermophiles (12). COG1857 is found in two SBCs (**1819.0** and **1819.1**) whose members combined include 22 proteins from hyperthermophiles and 4 proteins from thermophiles. In contrast, COG1688 is found in 4 root SBCs, suggesting that it includes proteins that are quite dissimilar. Nevertheless, it, too, is mostly associated with hyperthermophiles. Thus, our results, based on a larger number of sequenced Archaea, support the earlier prediction linking these COGs to thermophiles.

Oxygen usage: SBC **1278.0.0** contains 10 members, all annotated as oxygen-sensitive ribonucleoside-triphosphate reductases, from 10 organisms representing three phyla. This enzyme is known to be involved in anaerobic growth in some bacteria(13). Another ribonucleoside –related SBC (**1196.0**) contains members from the two major phyla and occur only in anaerobic organisms, thus suggesting a correlation in this case with aerobic lifestyle.

Metabolism: Interestingly enough, the only phylogenetically diverse SBCs here are from chemoorganoheterotrophs. Since most chemoorganoheterotrophs are thermophiles, it is difficult to separate the effects of thermal preference and metabolism. For instance, SBC **4682** contains pyrrolidone-carboxylate peptidases, a protein that is adapted for greater thermal stability in

hyperthermophiles (14) and that occurs equally among crenarchaea and euryarchaea. In general however, there are several SBCs related to sugar transport (**2078.0.0.1.0, 1924.1.0.0.1.0, 2078.0.1.0, 1924.1.0.0.0.1, 25.1.1.0.0.0.1.0.0, 2695**), which is to be expected for heterotrophic archaea since they must transport metabolites.

Habitat: Only 8 SBCs were selected here and the members of all 8 come from organisms assigned to aquatic or marine habitats. The function of most of their member proteins was annotated as *hypothetical*, but one SBC contains transposases (mobile elements by definition) and another contains threonyl-tRNA synthetases that were likely acquired from bacteria (15). This lack of any obvious candidates for proteins required for a particular habitat could be due to any of several factors: the number of proteins unique to and essential for a particular habitat is low, the overwhelming majority of such proteins are transmitted predominantly by vertical inheritance, and/or the habitat classifications are broad and include many microhabitats.

Correlating leaf clusters with phylogeny

To assess how closely the clustering paralleled phylogeny, we calculated the consistency scores for all leaf clusters based on the taxonomic assignments of their members. Figure 2 shows the distribution of the scores obtained for each of five taxonomic levels: phylum, class, order, family, and genus. The proportion of fully consistent clusters is greatest at the phylum level (76%, or 10,317of 13,586) and decreases with the decreasing phylogenetic distance, reaching a low of 19% (2,556 of 13,586) at the genus level. Thus, as expected, fewer protein classes were found to be unique to a genus than to a higher level such as a class or phylum.

Further interpretations can be derived from the observed consistency scores at the phylum level (c_P). However, additional factors need to be considered when interpreting inconsistent clusters. First, when a cluster partitions into two phyla, it might be because the member proteins are unique to one phylum or the other. However, it might instead be because the sequence conservation is relaxed enough that the evolutionary divergence of the phyla becomes the main factor driving the partitioning. Secondly, strongly conserved SBCs, such as those containing housekeeping proteins, would tend to show low consistency scores due to their ubiquity among archaea. Our data shows 3,066 conserved leaf SBCs where c_P is ≤ 0.8 and 10,317 fully consistent SBCs ($c_P = 1$). Thus, about 24% of archaeal proteins are conserved to a degree that overshadows the divergence of the archaeal phyla, or have been transferred.

LGT has been observed in a variety of organisms, but appears to be especially common in archaea

where 50% have acquired one or more protein domains from other organisms (16). Consistency scores can provide an indication of past LGT events. If an SBC acquired a sizeable portion of its members from another phylum, its c_P will be low—in the same range as for highly conserved protein families. However, if the transfer was recent and is reflected in only a very few members, consistency scores will be high (i.e., c_P will be close to 1). There are 203 such leaf SBCs in the data set, defined as clusters whose c_P is between 0.8 and 1. This is to be compared with 10,274 fully consistent SBCs ($c_P = 1$) and 2089 SBCs of conserved proteins ($c_P \leq 0.8$). Taken together, this suggests that approximately 1% of archaeal proteins have been recently transferred between phyla. The rate of transfer is doubtlessly much higher, since this value represents only transfers between phyla. At lower taxonomic levels, the number of LGT-candidate SBCs increases. However, these may be false positives because proteins in closely related organisms can be similar enough to cluster together without any LGT being involved.

Possible LGT could also be detected by identifying SBCs that are consistent with respect to functional annotation but inconsistent with respect to phylogeny. A simple example of this would be finding a functionally-consistent leaf SBC which contains mostly euryarchaeal proteins but also has one or two members from crenarchaeal organisms. We provide a few examples of this, including their corresponding SBC identifiers and locus tags for individual proteins, in the following paragraphs.

The anaerobic Crenarchaeota *Thermofilum pendens* Hrk 5 is the likely recipient of several transferred proteins, including some from the Euryarchaeota. In SBC **1229.0**, its FeoA family protein (Tpen_0974) is fully connected with 8 euryarchaeal ferrous iron transport proteins from organisms that are mostly anaerobic methane-producers. *T. pendens* also has a geranylgeranyl reductase (Tpen_0718) gene which is difficult to separate from euryarchaeal orthologs in SBC **8420.1.1.0**. Furthermore, SBC **870** suggests that it acquired a pyruvoyl-dependent arginine decarboxylase (Tpen_0872) from anaerobes that share its hot spring environment (17). This is more intriguing since most crenarchaeota do not have a recognizable arginine decarboxylase. Other possible transfers into *T. pendens* include a Rubredoxin-type Fe(Cys)4 protein (Tpen_1457) and a UDP-N-acetylglucosamine 2-epimerase (Tpen_1715), again from euryarchaeal classes such as the Methanobacteria, Methanococci, and Methanomicrobia (SBCs **8376.1.0.1.0.1.1.1.0.0.1.0.1.1.1.0.1** and **1255.1**).

SBCs **796** and **53.0.1** show that the crenarchaeon *Caldivirga maquilingensis* IC-167 may have obtained genes coding for diaminopimelate decarboxylase (Cmaq_0523) and thiamine-phosphate pyrophosphorylase (Cmaq_0061) from the Euryarchaeota. Likewise, SBC **553.1.0.1.0** is composed entirely of adenine deaminases from anaerobic species, all from Euryarchaeota except for the adenine

deaminase (Smar_0382) from the crenarchaeon *Staphylothermus marinus* F1.

In some cases, we instead found few Euryarchaeota among many Crenarchaeota. SBC **1306.1.1** contains crenarchaeal phosphohistidine phosphatases and one protein (AF1002) from the euryarchaeon *Archaeoglobus fulgidus* DSM 4304 which is annotated as hypothetical in IMG(4). All the taxa represented in this SBC are aquatic thermophiles or hyperthermophiles, thus suggesting that these organisms may at times share environments.

Frequent genetic input via LGT has been suggested for two groups of thermoacidophilic archaea: Sulfolobus and Thermoplasma (18). Since our sample includes three Sulfolobus and two Thermoplasma archaea, we would expect lower consistency scores ($c_P \leq 0.8$) to result from LGT in this case. Thus, contrary to what one might expect, rampant lateral transfer may be more difficult to detect. We found 30 SBCs with members from all five of these thermoacidophiles that also contain fewer than 10 proteins—thus avoiding ubiquitous conserved proteins. Although many of their members are of unknown function, they also include such diverse proteins as rieske-type iron-sulfur proteins, dihydrodipicolinate synthases, and cytochrome-related proteins.

Deep-branching archaea

Candidatus Korarchaeum cryptofilum(19) is an unculturable hyperthermophilic Archaeon isolated from hot springs and sediments, and is believed to represent a deep-branching phylum at the base of the two main archaeal phyla Crenarchaeota and Euryarchaeota. The gene content of *C* K cryptofilum is also largely hybrid, since it shares a majority of proteins with Crenarchaeota but also scattered components with Euryarchaeota; components often associated with replication/repair mechanisms. The authors suggest that these patterns could have arisen due to several lateral transfers. This can be difficult to determine by phylogenetic consistencies since we have only one representative of the Korarchaeota. For instance, *C* K cryptofilum shares a DNA polymerase II protein with euryarchaeal genomes (SBC **270**), suggesting either a strong conservation or lateral transfer. In SBC **1067**, lysyl tRNA is more distinct from the bulk of euryarchaeal lysyl-tRNA synthesases, therefore suggesting a vertical inheritance. A CO dehydrogenase maturation factor (SBC **4.0.0.1.0.1.0.0**) also seems inseparable from its euryarchaeal counterparts, which may suggest lateral transfer. NADH dehydrogenase subunit A (SBC **31.1.0**) however, clusters mainly with its crenarchaeal counterparts, although is slightly more dissimilar to the others. Furthermore, tRNA pseudouridine synthase (SBC **413**) belongs to a predominantly crenarchaeal cluster, albeit an outlier in the set.

The presence of *C* K cryptofilum in the dataset does not seem play the role of a natural breakpoint

between SBCs, i.e. its proteins are not usually intermediates in homology between Crenarchaeota and Euryarchaeota. For instance, SBCs **4.0.0.0.0.0.1.0.0.0.1** and **31.1** can be subpartitioned, but not by virtue of proteins belonging to *C* K cryptofilum. Another example is SBC **1122**, where the proteins that connect the two child clusters belong to a fused gene (Mbur_2244) from Methanococcoides burtonii DSM 6242, along with two proteins from the Thaumarchaeota rather than belonging to *C* K cryptofilum. Thus, it seems as if the genetic composition of *C* K cryptofilum has been influenced by LGT and gene loss, as suggested by Elkins and co-workers in the original paper(19).

Online tool

All SBCs are publicly available at http://coal.jgi-psf.org/. The search function provided enables you to locate clusters based on numerous parameters, including keywords, consistency scores, number of member proteins, and associated phenotypic metadata. Selecting a cluster displays a plot of the sequence orthology of its member proteins, thus visually communicating their relatedness. Furthermore, the user can interactively color the plot to indicate the functional annotation, phylogenetic assignment, or phenotypic metadata associated with each member protein. For example, Figure 3 shows the display for SBC **4** colored to indicate the taxonomic class of the source organisms. For each SBC, the interface also provides a functional summary, a representative member sequence, and convenient links to parent clusters and subclusters, where available. Proteins belonging to clusters can also be transferred to the IMG (4) gene cart for further analysis.

Discussion

We performed a topology-based soft clustering of the proteins from 53 archaeal genomes and evaluated the suitability of this method in light of the great variation in both sequence conservation and divergence effects between protein families. Integrating functional annotations, phylogeny, and associated phenotype with the sequence data allowed us to evaluate the influences of phylogeny and lifestyle on protein families. This methodology provides a valuable framework for biological data mining, of which we report some findings. Our clusters are publicly available for further exploration at http://coal.jgi-psf.org/

Materials and Methods

Data was collected from IMG (4) for all 53 currently available sequenced archaeal genomes (34 Euryarchaeota, 15 Crenarchaeota, two Thaumarchaeota, one Korarchaeota and one Nanoarchaeota) and bidirectional best hits were calculated between the genomes for all protein sequences using BLAST(20)

(blastp, default parameters and BLOSUM62 matrix). An initial low-end cutoff (the only hard cutoff in this study) was set at 30% protein identity and e-values < 10^{-6} as a reasonable limit of functional relevance. IMG proteins were assigned to COGs using reverse position-specific BLAST on position-specific scoring matrices provided by CDD(21).

Clustering

We have applied the spectral clustering procedure described previously (1,2) to the set of archaeal proteins. The proteins are represented as nodes in a connected undirected graph with edges that carry weights based on node-to-node similarity according to the protein identity. The clustering procedure is analogous to a random walk of a particle moving over the nodes of the graph. At each transition, the particle moves to an adjacent node with probabilities corresponding to the weights of the edges. The amount of time the particle spends in a given subgraph will determine whether this is indeed a cluster of its own or not.

Preliminary to clustering, our data was partitioned into disjoint block-matrices. Eigenvalues and eigenvectors were calculated using an ARPACK library specifically designed for large scale symmetric eigenproblems where the eigenvectors of the largest eigenvalues are desired.

The largest eigenvalue will equal one since all block matrices are stochastic (i.e., row sums to one). The magnitude of the second largest eigenvalue will determine how fast the corresponding Markov chain will approach its stationary distribution (22). In this approach, if the second largest eigenvalue is larger than 0.8, we further divide the block matrix into two new blocks using K-means clustering of eigenvectors 1 and 2. The eigenvalue cutoff of 0.8 therefore reflects the topology of the protein similarity graph, and not the actual similarity scores.

Metadata was added in the form of COGs(6), arCOGs(7), phenotype(5) and phylogeny(4). The phenotypic data was roughly categorized as habitat, metabolism, oxygen usage and preferred temperature range. For the habitat mappings, we intentionally used broad categories. For instance, *marine* is a catch-all phrase for e.g. deep sea vents and seawater. While this may seem overly insensitive in some cases, it is arguably the most manageable form of grouping habitats.

Consistency

The calculated consistency scores (see below) for each SBC reflect the degree of homogeneity of each type of metadata for that cluster. For instance, a cluster is said to be fully consistent with respect to phylogeny if all its member proteins belong to the same phylogenetic group. Similarly, a cluster would

be consistent with respect to its oxygen requirements if all member proteins belong to species that were annotated as *aerobic*.

Given that the consistency of an SBC (c) for a given metadata or phylogeny type (k) is inversely proportional to the degree of entropy (E), the consistency score c_k can be calculated as follows:

$$c_k = 1 - E_k = \sum_{i=1\ldots m} \frac{P(k_i)\ln(P(k_i))}{\ln(m)} + 1 \text{ for } m > 1$$

$$c_k = 1 \text{ for } m = 1$$

where k is the metadata or phylogeny type (e.g., *phylum*), i specifies each of the possible character states for k (e.g., *Crenarchaeota, Euryarchaeota*), $P(k_i)$ is the proportion of SBC member proteins with character state k_i, and m is the number of different character states observed in the SBC. Consistency scores are bounded by 0 and 1, with 1 being a fully consistent SBC. However, scoring was modified for the function metadata type when member proteins lacked annotation and thus could not be assigned to a COG. For these SBCs, the consistency score was defined as 0.

False positives

As with any clustering procedure, there is always a false positive rate. Potential errors may be from several sources such as the BLAST algorithm, biological complications such as fused genes and shared domains, and also from consistency calculations.

For biological error sources, the subpartitioning itself may serve as a mitigating factor. For instance, fused genes can often be a nuisance in clustering, since they tend to merge clusters which may otherwise not be related. However, this structure would serve as a natural breakpoint in the next round of partitioning, so once again the two clusters would be distinct. Shared domains may also behave analogously to fused genes by bringing together protein groups otherwise dissimilar. Again, this would very likely be resolved in the next round of partitioning.

For the consistency measure, the false positive rate depends on the number of protein members and the number of characters. We simulated 10,000 SBCs of 10 proteins, with 5 characters occurring at an equal rate. We found that under these conditions the proportion of SBCs with a consistency score c of > 0.7 was very low (10^{-4}). When one character is more dominant, such as for phyla where 34 genomes are of the same character, it is far easier to find fully consistent SBCs by chance. Through simulation, we found that 43% of the SBCs were fully consistent (c=1) as compared to the observed 76%. However, in case the significance of the phylogenetic consistencies is questioned, it is easy to quality

check the measure by studying the next level of phylogeny. At the class level, the expected frequency of fully consistent SBCs drops to 1.1% compared to the observed frequency 59%. For some phenotypic characters, false positives may still be a problem. For instance, 39% of SBCs are consistent with oxygen usage, whereas we expect 15% by chance. While fully consistent SBCs are significantly overrepresented, we still suggest that the protein annotation be consulted as a quality check to see if indeed the functions may be relevant to the oxygen usage of the organism.

Funding

This work was supported by and performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

Acknowledgments

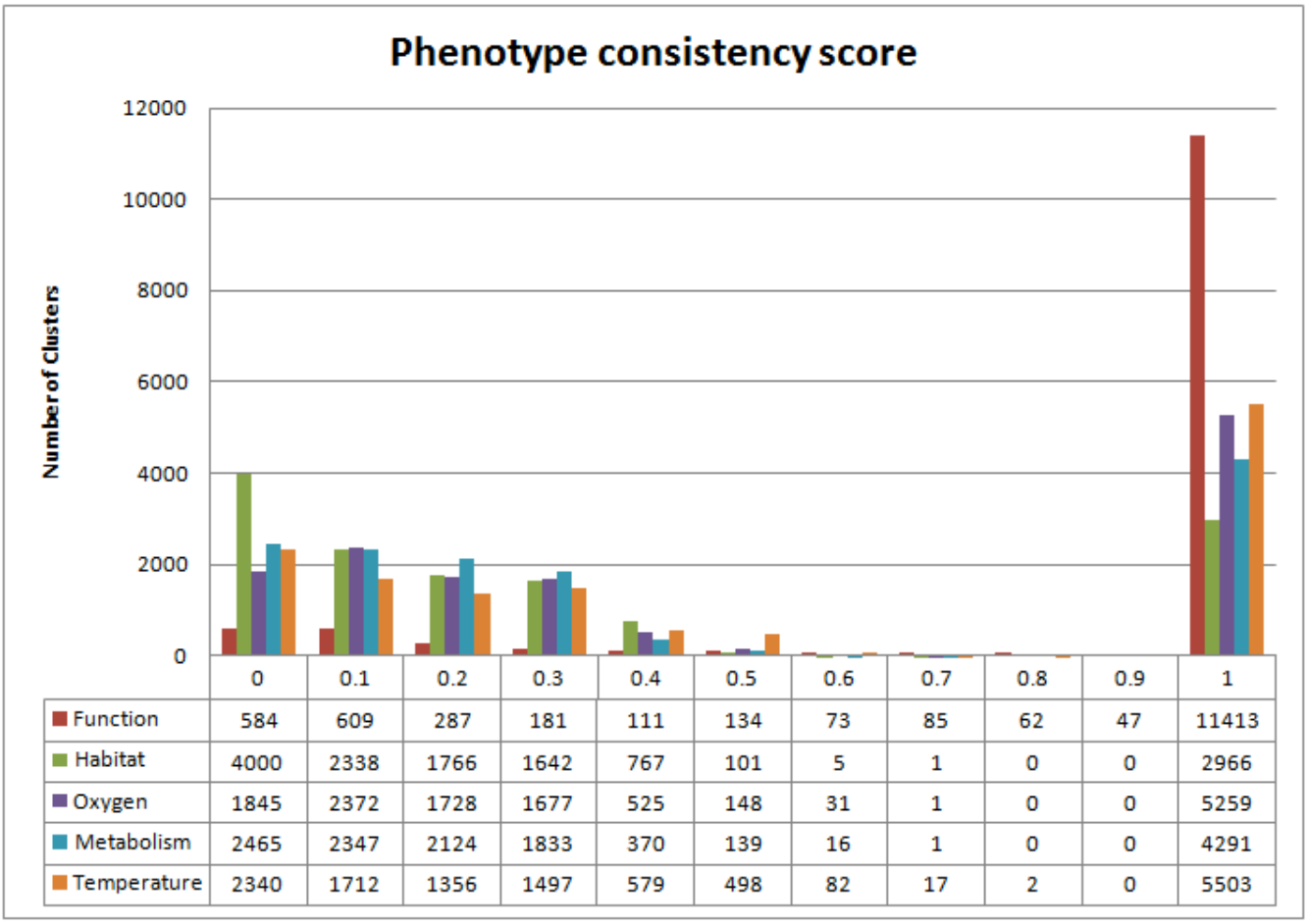
The authors wish to thank N. Ivanova and A. Lykidis for constructive discussion, K. Palaniappan and E. Szeto for their technical expertise and L. Looney for graphics used in the web page.

References

1. Paccanaro, A., Casbon, J.A. and Saqi, M.A. (2006) Spectral clustering of protein sequences. *Nucleic Acids Res*, **34**, 1571-1580.
2. Brewer, M.L. (2007) Development of a spectral clustering method for the analysis of molecular data sets. *J Chem Inf Model*, **47**, 1727-1733.
3. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**, 4285-4288.
4. Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.M., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K. *et al.* (2007) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res*, **36**, D528-533.
5. Liolios, K., Mavromatis, K., Tavernarakis, N. and Kyrpides, N.C. (2007) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*.
6. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
7. Makarova, K.S., Sorokin, A.V., Novichkov, P.S., Wolf, Y.I. and Koonin, E.V. (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct*, **2**, 33.
8. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res*, **32**, D138-141.
9. Kumagai, I., Watanabe, K. and Oshima, T. (1980) Thermally induced biosynthesis of 2'-O-methylguanosine in tRNA from an extreme thermophile, Thermus thermophilus HB27. *Proc Natl Acad Sci U S A*, **77**, 1922-1926.
10. Forterre, P. (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet*, **18**, 236-237.
11. Brochier-Armanet, C. and Forterre, P. (2007) Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. *Archaea*, **2**, 83-93.
12. Makarova, K.S., Aravind, L., Grishin, N.V., Rogozin, I.B. and Koonin, E.V. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res*, **30**, 482-496.
13. Fontecave, M., Eliasson, R. and Reichard, P. (1989) Oxygen-sensitive ribonucleoside triphosphate reductase is present in anaerobic Escherichia coli. *Proc Natl Acad Sci U S A*, **86**, 2147-2151.
14. Ogasahara, K., Khechinashvili, N.N., Nakamura, M., Yoshimoto, T. and Yutani, K. (2001) Thermal stability of pyrrolidone carboxyl peptidases from the hyperthermophilic Archaeon, Pyrococcus furiosus. *Eur J Biochem*, **268**, 3233-3242.
15. Rigden, D.J. (2004) Archaea recruited D-Tyr-tRNATyr deacylase for editing in Thr-tRNA synthetase. *Rna*, **10**, 1845-1851.
16. Choi, I.G. and Kim, S.H. (2007) Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A*, **104**, 4489-4494.
17. Hetzer, A., Morgan, H.W., McDonald, I.R. and Daughney, C.J. (2007) Microbial life in Champagne Pool, a geothermal spring in Waiotapu, New Zealand. *Extremophiles*, **11**, 605-614.
18. Angelov, A. and Liebl, W. (2006) Insights into extreme thermoacidophily based on genome analysis of Picrophilus torridus and other thermoacidophilic archaea. *J Biotechnol*, **126**, 3-10.
19. Elkins, J.G., Podar, M., Graham, D.E., Makarova, K.S., Wolf, Y., Randau, L., Hedlund, B.P., Brochier-Armanet, C., Kunin, V., Anderson, I. *et al.* (2008) A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci U S A*, **105**, 8102-8107.
20. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
21. Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D. *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res*, **35**, D237-240.

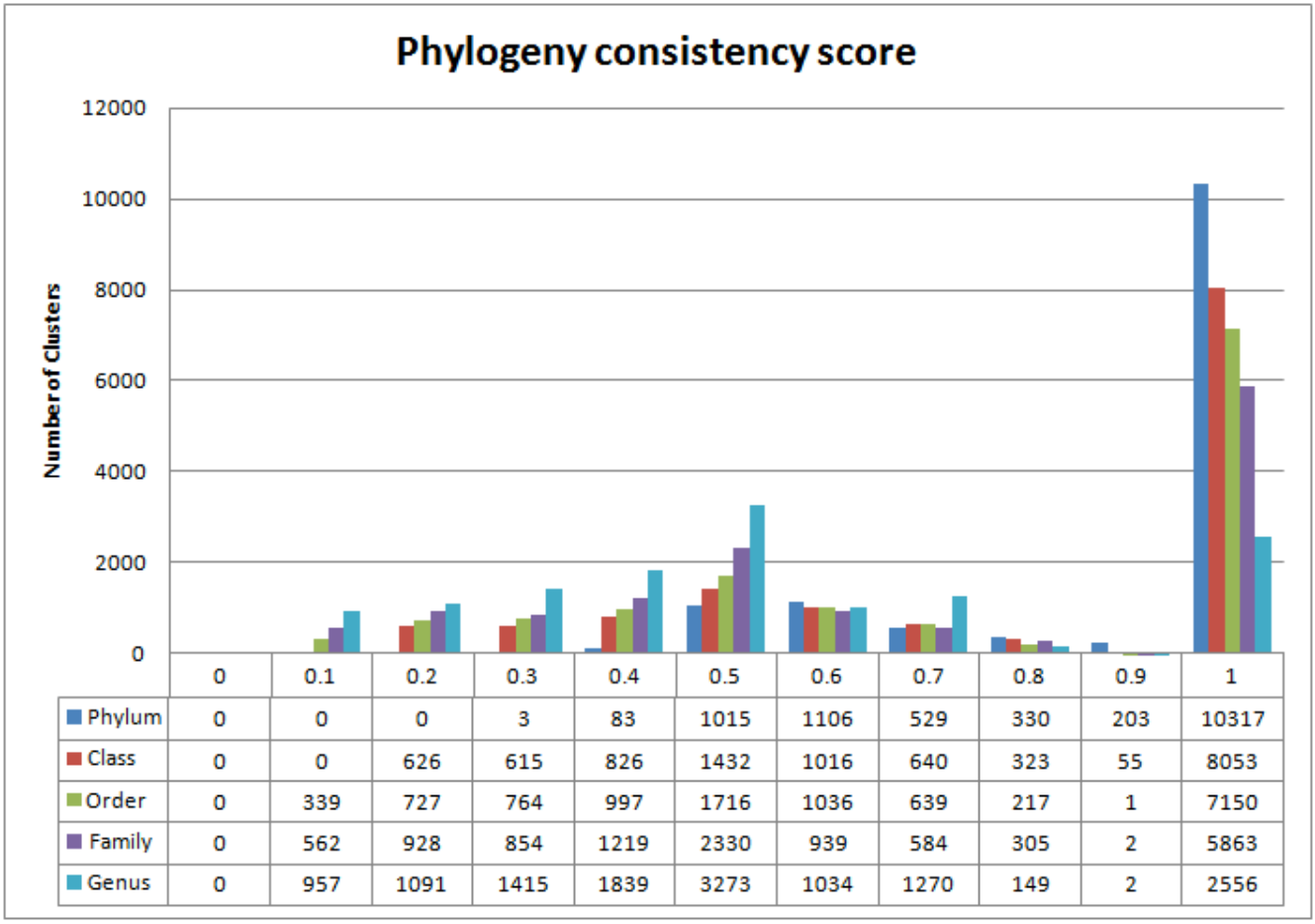
22. *Broder, A. and Karlin, A. (1988) Bounds on the cover time. Journal of Theoretical Probability, **2**, 101-120.*

Figures



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Function	584	609	287	181	111	134	73	85	62	47	11413
Habitat	4000	2338	1766	1642	767	101	5	1	0	0	2966
Oxygen	1845	2372	1728	1677	525	148	31	1	0	0	5259
Metabolism	2465	2347	2124	1833	370	139	16	1	0	0	4291
Temperature	2340	1712	1356	1497	579	498	82	17	2	0	5503

Fig 1 Distribution of leaf cluster consistency scores for functional annotation and four phenotypic metadata types. A score of 1 indicates that all proteins in the cluster share the same character state (e.g., all belong to hyperthermophiles). A score of <1 means that the cluster includes member proteins with different character states (e.g., some from hyperthermophiles and some from mesophiles). If any cluster member lacked a functional annotation, we defined that cluster's consistency score for function metadata as 0.



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Phylum	0	0	0	3	83	1015	1106	529	330	203	10317
Class	0	0	626	615	826	1432	1016	640	323	55	8053
Order	0	339	727	764	997	1716	1036	639	217	1	7150
Family	0	562	928	854	1219	2330	939	584	305	2	5863
Genus	0	957	1091	1415	1839	3273	1034	1270	149	2	2556

Fig 2 Distribution of cluster consistency scores for phylogeny at each of five taxonomic levels.

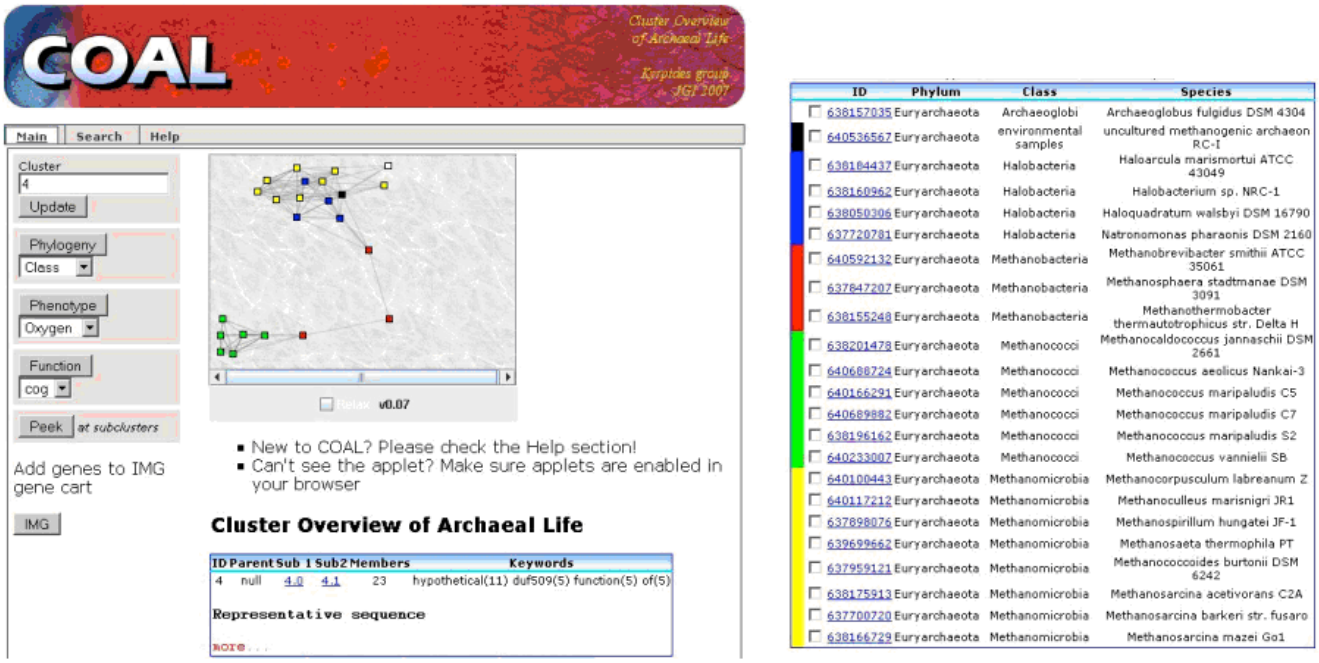


Fig 3 Screenshot of the COAL user interface available at http://coal.jgi-psf.org/

Tables

Table 1. The distribution of the number (*N*) of COGs, arCOGs and Pfams associated with individual SBCs. The members of most SBCs are all assigned to the same functional cluster (i.e. COG, arCOG, Pfam), but some SBCs contain members from for instance several COGs, and the members of approximately 4000 SBCs have no associated COGs. (Leaf SBCs also include those root SBCs that could not be partitioned.)

	Root SBC			Leaf SBC		
N	*COG*	*arCOG*	*Pfam*	*COG*	*arCOG*	*Pfam*
0	4248	5144	4212	4984	6643	4934
1	3682	2576	3373	7512	5920	6950
2	367	466	616	881	857	1342
3	100	110	156	157	121	271
4	24	59	52	34	32	67
5	14	32	21	15	9	13
6	8	16	11	1	3	4
7	2	14	6	2	0	4
8	2	9	3	0	1	1
9	5	10	3	0	0	0
More	11	27	10	0	0	0